

Low Rank Approximation and Regression in Input Sparsity Time

Kenneth L. Clarkson
IBM Almaden

David P. Woodruff
IBM Almaden

July 27, 2012

Abstract

We design a new distribution over $\text{poly}(r\epsilon^{-1}) \times n$ matrices S so that for any fixed $n \times d$ matrix A of rank r , with probability at least $9/10$, $\|SAx\|_2 = (1 \pm \epsilon)\|Ax\|_2$ simultaneously for all $x \in \mathbb{R}^d$. Such a matrix S is called a *subspace embedding*. Furthermore, SA can be computed in $\text{nnz}(A) + \text{poly}(d\epsilon^{-1})$ time, where $\text{nnz}(A)$ is the number of non-zero entries of A . This improves over all previous subspace embeddings, which required at least $\Omega(nd \log d)$ time to achieve this property. We call our matrices S *sparse embedding matrices*.

Using our sparse embedding matrices, we obtain the fastest known algorithms for overconstrained least-squares regression, low-rank approximation, and approximating all leverage scores.

- to output an x' for which $\|Ax' - b\|_2 \leq (1 + \epsilon) \min_x \|Ax - b\|_2$ for an $n \times d$ matrix A and an $n \times 1$ column vector b , we obtain an algorithm running in $O(\text{nnz}(A)) + \text{poly}(d\epsilon^{-1})$ time.
- to obtain a decomposition of an $n \times n$ matrix A into a product of an $n \times k$ matrix L , a $k \times k$ diagonal matrix D , and a $k \times n$ matrix R , for which

$$\|A - L \cdot D \cdot R\|_F \leq (1 + \epsilon)\|A - A_k\|_F,$$

where A_k is the best rank- k approximation, our algorithm runs in $O(\text{nnz}(A)) + n \cdot \text{poly}(k\epsilon^{-1})$ time.

- to output an approximation to all leverage scores of an $n \times d$ input matrix A simultaneously, with constant relative error, our algorithms run in $O(\text{nnz}(A) \log n) + \text{poly}(r \log n)$ time.

We optimize the polynomial factors in the above stated running times, and show various tradeoffs. We also provide preliminary experimental results which suggest that our algorithm is competitive in practice.

1 Introduction

A large body of work has been devoted to the study of fast randomized approximation algorithms for problems in numerical linear algebra. Several well-studied problems in this area include least squares regression, low rank approximation, and approximate computation of leverage scores. These problems have many applications in data mining [5], recommendation systems [17], information retrieval [37], web search [1, 32], clustering [13, 34], and learning mixtures of distributions [31, 2]. The use of randomization and approximation allows one to solve these problems much faster than deterministic methods.

For example, in the overconstrained least-squares regression problem, we are given an $n \times d$ matrix A of rank r as input, $n \gg d$, together with an $n \times 1$ column vector b . The goal is to output a vector x' so that with high probability, $\|Ax' - b\|_2 \leq (1 + \epsilon) \min_x \|Ax - b\|_2$. The minimizing vector x^* can be expressed in terms of the Moore-Penrose pseudoinverse A^+ of A , namely, $x^* = A^+b$. If A has full column rank, this simplifies to $x^* = (A^T A)^{-1} A^T b$. This minimizer can be computed deterministically in $O(nd^2)$ time, but with randomization and approximation, this problem can be solved in $O(nd \log d) + \text{poly}(d(\ln n)/\epsilon)$ time [39, 24], which is much faster for $d \ll n$ and ϵ not too small.

Another example is low rank approximation. Here we are given an $n \times n$ matrix (which can be generalized to $n \times d$) and an input parameter k , and the goal is to find an $n \times n$ matrix A' of rank at most k for which $\|A' - A\|_F \leq (1 + \epsilon)\|A - A_k\|_F$, where for an $n \times n$ matrix B , $\|B\|_F^2 \equiv \sum_{i=1}^n \sum_{j=1}^n B_{i,j}^2$ is the squared

Frobenius norm, and $A_k \equiv \operatorname{argmin}_{\operatorname{rank} B \leq k} \|A - B\|_F$. Here A_k can be computed deterministically using the singular value decomposition in $O(n^3)$ time. However, using randomization and approximation, this problem can be solved in $O(\operatorname{nnz}(A) \cdot (k/\varepsilon + k \log k) + n \cdot \operatorname{poly}(k/\varepsilon))$ time [39, 8], where $\operatorname{nnz}(A)$ denotes the number of non-zero entries of A . The problem can also be solved using randomization and approximation in $O(n^2 \log n) + n \cdot \operatorname{poly}(k/\varepsilon)$ time [39], which may be faster than the former for dense matrices and large k .

Another problem we consider is approximating the *leverage scores*. Given an $n \times d$ matrix A with $n \gg d$, one can write $A = U\Sigma V^\top$ in its singular value decomposition, where the columns of U are the left singular vectors, Σ is a diagonal matrix, and the columns of V are the right singular vectors. Although U has orthonormal columns, not much can be immediately said about the squared lengths $\|U_i\|_2^2$ of its rows. These values are known as the leverage scores, and measure the extent to which the singular vectors of A are correlated with the standard basis. They are basis-independent, that is, the leverage scores of a matrix A are equal to the diagonal elements of the projection matrix onto the span of the columns of A ; see [19] for background on leverage scores as well as a list of applications. The leverage scores will also play a crucial role in our work, as we shall see. The goal of approximating the leverage scores is to, simultaneously for each $i \in [n]$, output a constant factor approximation to $\|U_i\|_2^2$. Using randomization, this can be solved in $O(nd \log n + d^3 \log d \log n)$ time [19].

There are also solutions for these problems based on sampling. They either get a weaker additive error [25, 37, 3, 14, 15, 16, 20, 38, 11], or they get bounded relative error but are slow [12, 21, 22, 23]. Many of the latter algorithms were improved independently by Deshpande and Vempala [12] and Sarlós [39], and in followup work [24, 36, 33]. There are also solutions based on iterative and conjugate-gradient methods, see, e.g., [41], or [42] as a recent example. These methods repeatedly compute matrix-vector products Ax for various vectors x ; in the most common setting, such products require $\Theta(\operatorname{nnz}(A))$ time. Thus the work per iteration of these methods is $\Theta(\operatorname{nnz}(A))$, and the number of iterations N that are performed depends on the desired accuracy, spectral properties of A , numerical stability issues, and other concerns, and can be large. A recent survey suggests that N is typically $\Theta(k)$ for using Krylov methods (such as Arnoldi and Lanczos iterations) to approximate the k leading singular vectors [26]. One can also use some of these techniques together, for example by first obtaining a preconditioner using the Johnson-Lindenstrauss (JL) transform, and then running an iterative method.

While these results illustrate the power of randomization and approximation, their main drawback is that they are not optimal. For example, for regression, ideally we could hope for $O(\operatorname{nnz}(A)) + \operatorname{poly}(d/\varepsilon)$ time. While the $O(nd \log d) + \operatorname{poly}(d/\varepsilon)$ time algorithm for least squares regression is almost optimal for *dense* matrices, if $\operatorname{nnz}(A) \ll nd$, say $\operatorname{nnz}(A) = O(n)$, as commonly occurs, this could be much worse than an $O(\operatorname{nnz}(A)) + \operatorname{poly}(d/\varepsilon)$ time algorithm. Similarly, for low rank approximation, the best known algorithms that are condition-independent run in $O(\operatorname{nnz}(A)(k/\varepsilon + k \log k) + n \cdot \operatorname{poly}(k/\varepsilon))$ time, while we could hope for $O(\operatorname{nnz}(A)) + \operatorname{poly}(k/\varepsilon)$ time.

1.1 Results

We resolve the above gaps by achieving algorithms for least squares regression, low rank approximation, and approximate leverage scores, whose leading order term in the time complexity is $O(\operatorname{nnz}(A))$, with a constant factor that is independent of any properties of A . Our results are as follows:

- **Least Squares Regression:** We present two different algorithms for an $n \times d$ matrix A with rank r and given $\varepsilon > 0$. The first has running time $O(\operatorname{nnz}(A) + d^5 \varepsilon^{-4} \log^3(d/\varepsilon))$. The second has running time $O(\operatorname{nnz}(A) \log(d/\varepsilon) + r^3 \varepsilon^{-2} (\log(r/\varepsilon)) (\log(1/\varepsilon)))$.
- **Low Rank Approximation:** We achieve $O(\operatorname{nnz}(A)) + n \cdot \operatorname{poly}(k/\varepsilon)$ time. More specifically, as stated by Theorem 31, for k small enough, we obtain $O(\operatorname{nnz}(A) + nk^4 \varepsilon^{-8} \log^2(k/\varepsilon))$ time, which is $O(\operatorname{nnz}(A) + nk^4 \log^2 k)$ for fixed ε , or $O(\operatorname{nnz}(A) \log^2(k/\varepsilon) + nk \varepsilon^{-8} \log(k/\varepsilon)(k + \log(1/\varepsilon)))$ time, which is $O(\operatorname{nnz}(A) \log^2 k + nk^2 \log k)$ for fixed ε . Here $k \leq n^{1/2-\gamma}$ is small enough, for any fixed $\gamma > 0$; we also have results for larger k . As described for Theorem 30, we also obtain an approximation to A of rank $O(k/\varepsilon)$ in time $O(\operatorname{nnz}(A) + \operatorname{poly}(k/\varepsilon))$, which can be an improvement in running time even when $\operatorname{nnz}(A)/n$ is not large; the approximation is represented implicitly, as the product of three matrices.

- **Approximate Leverage Scores:** For any fixed constant $\varepsilon > 0$, we simultaneously approximate all n leverage scores in $O(\text{nnz}(A) \log n + r^3 \log^2 r + r^2 \log n)$ time. This can be generalized to sub-constant ε to achieve $O(\text{nnz}(A) \log n) + \text{poly}(r/\varepsilon)$ time, though in the applications we are aware of, such as coresets for regression [9], ε is typically constant (in the applications of this, a general $\varepsilon > 0$ can be achieved by over-sampling [21, 9]).

1.2 Techniques

All of our results are achieved by improving the time complexity of computing what is known as a *subspace embedding*. For a given $n \times d$ matrix A , call $S : \mathbb{R}^n \mapsto \mathbb{R}^t$ a *subspace embedding matrix* for A if, for all $x \in \mathbb{R}^d$, $\|SAx\|_2 = (1 \pm \varepsilon)\|Ax\|_2$. That is, S embeds the column space $C(A) \equiv \{Ax \mid x \in \mathbb{R}^d\}$ into \mathbb{R}^t while approximately preserving the norms of all vectors in that subspace.

The *subspace embedding problem* is to find such an embedding matrix obviously, that is, to design a distribution π over linear maps $S : \mathbb{R}^n \mapsto \mathbb{R}^t$ such that for any fixed $n \times d$ matrix A , if we choose $S \sim \pi$ then with large probability, S is an embedding matrix for A . The goal is to minimize t as a function of n, d , and ε , while also allowing the matrix-matrix product $S \cdot A$ to be computed quickly.

By taking S to be a Fast Johnson Lindenstrauss transform, one can set $t = O(d/\varepsilon^2)$ and achieve $O(nd \log t)$ time for $d < n^{1/2-\gamma}$ for any constant $\gamma > 0$. One can also take S to be a subsampled randomized Hadamard transform (see, e.g., [24]) and set $t = O(d \ln n (\ln d + \ln \ln d + \ln 1/\varepsilon)/\varepsilon^2)$, to achieve $O(nd \log t)$ time. These were the fastest known subspace embeddings achieving any value of t not depending polynomially on n . Our main result improves this to achieve $t = \text{poly}(d/\varepsilon)$ for matrices S for which SA can be computed in $\text{nnz}(A)$ time! Given our new subspace embedding, we plug it into known methods of solving the above linear algebra problems given a subspace embedding as a black box.

In fact, our subspace embedding is nothing other than the **CountSketch** matrix in the data stream literature [6], see also [40]. This matrix was also studied by Dasgupta, Kumar, and Sarlós [10]. Formally, S has a single randomly chosen non-zero entry $S_{h(j),j}$ in each column j , for a random mapping $h : [n] \mapsto [t]$. With probability $1/2$, $S_{h(j),j} = 1$, and with probability $1/2$, $S_{h(j),j} = -1$.

While such matrices S have been studied before, the surprising fact is that they actually provide subspace embeddings. Indeed, the usual way of proving that a random $S \sim \pi$ is a subspace embedding is to show that for any fixed vector $y \in \mathbb{R}^d$, $\Pr[\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2] \geq 1 - \exp(-d)$. One then puts a net (see, e.g., [4]) on the unit vectors in the column space $C(A)$, and argues by a union bound that $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ for all net points y . Then, since the net is sufficiently fine, and since the mapping is linear, this implies that $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ for all vectors $y \in C(A)$.

We stress that our choice of matrices S does not preserve norms of an arbitrary set of $\exp(d)$ vectors with high probability, and so the above approach cannot work for our choice of matrices S . We instead critically use that these $\exp(d)$ vectors all come from a d -dimensional subspace (namely, $C(A)$), and therefore have a very special structure. The structural fact we use is that there is a fixed set H of size d/α which depends only on the subspace, such that for any unit vector y in the subspace, H contains the indices of all coordinates of y larger than $\sqrt{\alpha}$ in magnitude. The key property here is that the set H is independent of y , or in other words, only a small set of coordinates could ever be large as we range over all unit vectors in the subspace. The set H is exactly the set of large leverage scores of the column space of A !

Given this observation, by setting $t \geq K|H|^2$ for a large enough constant K , we have that with probability $1 - 1/K$, there are no two distinct $j \neq j'$ with $j, j' \in H$ for which $h(j) = h(j')$. That is, the coordinates in H are “perfectly hashed” with large probability. Call this event \mathcal{E} , which we condition on.

Given a unit vector y in the subspace, we can write it as $y^H + y^L$, where y^H consists of y with the coordinates in $[n] \setminus H$ replaced with 0, while y^L consists of y with the coordinates in H replaced with 0. We seek to bound

$$\|Sy\|_2^2 = \|Sy^H\|_2^2 + \|Sy^L\|_2^2 + 2\langle Sy^H, Sy^L \rangle.$$

Since \mathcal{E} occurs, we have the isometry $\|Sy^H\|_2^2 = \|y^H\|_2^2$. Now, $\|y^L\|_\infty^2 < \alpha$, and so we can apply Theorem 2 of [10] which shows that for mappings of our form, if the input vector has small infinity norm, then S

preserves the norm of the vector up to an additive $O(\varepsilon)$ factor with high probability. Here, it suffices to set $\alpha = 1/\text{poly}(d/\varepsilon)$.

Finally, we can bound $\langle Sy^H, Sy^L \rangle$ as follows. Define $G \subseteq [n] \setminus H$ to be the set of coordinates j for which $h(j) = h(j')$ for a coordinate $j' \in H$, that is, those coordinates in $[n] \setminus H$ which “collide” with an element of H . Then, $\langle Sy^H, Sy^L \rangle = \langle Sy^H, Sy^{L'} \rangle$, where $y^{L'}$ is a vector which agrees with y^L on coordinates $j \in G$, and is 0 on the remaining coordinates. By Cauchy-Schwarz, this is at most $\|Sy^H\|_2 \cdot \|Sy^{L'}\|_2$. We have already argued that $\|Sy^H\|_2 = \|y^H\|_2 \leq 1$ for unit vectors y . Moreover, we can again apply Theorem 2 of [10] to bound $\|Sy^{L'}\|_2$, since, conditioned on the coordinates of $y^{L'}$ hashing to the set of items that the coordinates of y^H hash to, they are otherwise random, and so we again have a mapping of our form (with a smaller t and applied to a smaller n) applied to a vector with small infinity-norm. Therefore, $\|Sy^{L'}\|_2 \leq O(\varepsilon) + \|y^{L'}\|_2$ with high probability. Finally, by Bernstein bounds, since the coordinates of y^L are small and t is sufficiently large, $\|y^{L'}\|_2 \leq \varepsilon$ with high probability. Hence, conditioned on event \mathcal{E} , $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ with probability $1 - \exp(-d)$, and we can complete the argument by union-bounding over a sufficiently fine net.

We note that an inspiration for this work comes from work on estimating norms in a data stream with efficient update time by designing separate data structures for the heavy and the light components of a vector [35, 30]. A key concept here is to characterize the heaviness of coordinates in a vector space in terms of its leverage scores.

Optimizing the additive term: The above approach already illustrates the main idea behind our subspace embedding, showing that it can be implemented in $\text{nnz}(A)$ time. This is sufficient to achieve our numerical linear algebra results in time $O(\text{nnz}(A)) + \text{poly}(d/\varepsilon)$ for regression and $O(\text{nnz}(A)) + n \cdot \text{poly}(k/\varepsilon)$ for low rank approximation. However, for some applications d, k , or $1/\varepsilon$ may also be large, and so it is important to achieve a small degree in the additive $\text{poly}(d/\varepsilon)$ and $n \cdot \text{poly}(k/\varepsilon)$ factors. The number of rows of the matrix S is $t = \text{poly}(d/\varepsilon)$, and the simplest analysis described above would give roughly $t = (d/\varepsilon)^8$. We now show how to optimize this.

The first idea for bringing this down is that the analysis of [10] can itself be tightened by using that we are applying it on vectors coming from a subspace instead of on a set of arbitrary vectors. This involves observing that in the analysis of [10], if on input vector y and for every $i \in [t]$, $\sum_{j|h(j)=i} y_j^2$ is small then the remainder of the analysis of [10] does not require that $\|y\|_\infty$ be small. Since our vectors come from a subspace, it suffices to show that for every $i \in [t]$, $\sum_{j|h(j)=i} \|U_j\|_2^2$ is small, where $\|U_j\|_2^2$ is the j -th leverage score of A . Therefore we do not need to perform this analysis for each y , but can condition on a single event, and this effectively allows us to increase α in the outline above, thereby reducing the size of T , and also the size of t since we must have $t = \Omega(|H|^2)$. In fact, we instead follow a simpler and slightly tighter analysis of [29] based on the Hanson-Wright inequality.

The second idea is that the estimation of $\|y^H\|_2$, the contribution from the “heavy coordinates”, is inefficient since it requires a perfect hashing of the coordinates. Indeed, while the above optimization is tuned for efficiently estimating $\|y^L\|_2$, the number t of rows needed is still roughly $(d/\varepsilon)^4$ due to the estimation of $\|y^H\|_2$. We instead estimate the contribution from the heavy coordinates in a different manner. By standard balls-and-bins analyses, if we have $O(d^2/\log d)$ bins and d^2 balls, then with high probability each bin will contain $O(\log d)$ balls. We thus make t roughly d^2 and think of having $O(d^2/\log d)$ bins. In each bin i , $O(\log d)$ heavy coordinates j will satisfy $h(j) = i$. Then, we apply a separate JL transform on the coordinates that hash to each bin i . This JL transform maps a vector $z \in \mathbb{R}^n$ to an $O((\log d)/\varepsilon^2)$ -dimensional vector z' for which $\|z'\|_2 = (1 \pm \varepsilon)\|z\|_2$ with probability at least $1 - 1/\text{poly}(d)$. Since there are only $O(\log d)$ heavy coordinates mapping to a given bin, we can put a net on all vectors on such coordinates of size only $\text{poly}(d)$. We can do this for each of the $O(d^2/\log d)$ bins and take a union bound. It follows that the 2-norm of the vector of coordinates that hash to each bin is preserved, and so the entire vector y^H of heavy coordinates has its 2-norm preserved. By a result of [29], the JL transform can be implemented in $O((\log d)/\varepsilon)$ time, giving total time $O(\text{nnz}(A) \log d/\varepsilon)$, and this reduces t to roughly d^2/ε^4 .

We also note that for applications such as least squares regression, it suffices to set ε to be a constant in the subspace embedding, since we can use an approach in [21, 9] which, given constant-factor approximations to all of the leverage scores, can then achieve a $(1 + \varepsilon)$ -approximation to least squares regression by slightly

over-sampling rows of the adjoined matrix $A \circ b$ proportional to its leverage scores, and solving the induced subproblem. This results in a better dependence on ε .

We can also compose our subspace embedding with a fast JL transform to further reduce t to the optimal value of about d/ε^2 . Since $S \cdot A$ already has small dimensions, applying a fast JL transform is now efficient.

Finally, we can use a recent result of [7] to replace most dependencies on d in our running times for regression with a dependence on the rank r of A , which may be smaller.

1.3 Experiments

In §8, we give some preliminary experimental results for an application to low-rank approximation. Here we consider the simplest version of our techniques, where for sparse embedding matrices S and R , we compare the error of the low-rank approximation $AR^\top(SAR^\top)^-SA$ to A with that of the best rank- k approximation, for different sketching dimensions for S and R , and hence ranks for the low-rank approximation. The results are encouraging, and (as often happens) better than the worst-case bounds we are able to prove. We consider a broad class of matrices, and find, for example, that when $t/k \geq 4$, all but a few such approximations have error at most 1.21 times the error of the best rank- k matrix.

2 Sparse Embedding Matrices

Let $A \in \mathbb{R}^{n \times d}$. We assume $n > d$. Let $\text{nnz}(A)$ denote the number of non-zero entries of A . We can assume $\text{nnz}(A) \geq n$ and that there are no all-zero rows or columns in A .

For a parameter t , we define a random linear map $\Phi D : \mathbb{R}^n \rightarrow \mathbb{R}^t$ as follows:

- $h : [n] \mapsto [t]$ is a random map so that for each $i \in [n]$, $h(i) = t'$ for $t' \in [t]$ with probability $1/t$.
- $\Phi \in \{0, 1\}^{t \times n}$ is a $t \times n$ binary matrix with $\Phi_{h(i), i} = 1$, and all remaining entries 0.
- D is an $n \times n$ random diagonal matrix, with each diagonal entry independently chosen to be $+1$ or -1 with equal probability.

We will refer to matrices of the form ΦD as a *sparse embedding matrix*.

3 Analysis

Let $U \in \mathbb{R}^{n \times r}$ have columns that form an orthonormal basis for the column space $C(A)$. Let $U_{1,*}, \dots, U_{n,*}$ be the rows of U , and let $u_i \equiv \|U_{i,*}\|^2$. Note that $\sum_i u_i = \|U\|_F^2 = r$. A unit vector $y \in C(A)$ can be expressed as $y = Ux$ for unit vector x , and so $y_i^2 = (U_{i,*}x)^2 \leq u_i \|x\|^2 = u_i$.

Unless otherwise indicated, a vector norm $\|y\|$ is the ℓ_2 norm.

Let $T > 0$ be a parameter.

3.1 Handling vectors with small entries

We begin the analysis by considering fixed unit vectors $y \in C(A)$ such that for all $i \in [n]$ for which $u_i > T$, we have $y_i = 0$. Notice that since $y_i^2 \leq u_i$, this in particular implies that $\|y\|_\infty^2 \leq T$. We extend this to all unit vectors in subsequent sections.

The following is similar to Lemma 6 of [10], and is a standard balls-and-bins analysis.

Lemma 1 *For $T > 0$, $\delta \in (0, 1)$, and $t \geq Mr^2$ for M larger than a sufficiently large positive constant, let \mathcal{E}_h denote the event that*

$$\frac{2}{Mr} \geq \max_{j \in [t]} \sum_{\substack{i \in h^{-1}(j) \\ u_i \leq T}} u_i.$$

If $T \leq \frac{1}{6Mr \log(t/\delta)}$, then $\Pr[\mathcal{E}_h] \geq 1 - \delta$.

Proof: We will prove that the bound holds for fixed $j \in [t]$ with failure probability δ/t , and then apply a union bound.

Let X_i denote the random variable $u_i I_{h(i)=j, u_i \leq T}$. We will apply Bernstein's inequality to $\sum_i Y_i$ where $Y_i \equiv X_i - \mathbf{E}[X_i]$. We have $|Y_i| \leq T$, and with that bound, Bernstein's inequality states that

$$\log \Pr \left[\sum_{i \in [t]} Y_i \geq z \right] \leq \frac{-z^2/2}{\sum_i \mathbf{E}[Y_i^2] + zT/3}. \quad (1)$$

We have $\mathbf{E}[Y_i^2] \leq \mathbf{E}[X_i^2] \leq u_i T/t$ so that $\sum_i \mathbf{E}[Y_i^2] \leq rT/t$. Hence, (1) becomes

$$\log \Pr \left[\sum_{i \in [t]} Y_i \geq z \right] \leq \frac{-z^2/2}{rT/t + zT/3}. \quad (2)$$

Using that $T \leq 1/(6Mr \log(t/\delta))$ and that $t \geq Mr^2$, we have that $rT/t \leq 1/(6M^2r^2 \log(t/\delta))$. We set $z = 1/(Mr)$, and consequently have that $zT/3 \leq 1/(3M^2r^2 \log(t/\delta))$. Hence,

$$\log \Pr \left[\sum_{i \in [t]} Y_i \geq z \right] \leq \frac{-1/(2M^2r^2)}{1/(2M^2r^2 \log(t/\delta))} = -\log(t/\delta).$$

Now since $\sum_i \mathbf{E}[X_i] \leq r/t \leq 1/(Mr)$,

$$\Pr \left[\sum_{i \in [t]} X_i \geq \frac{2}{Mr} \right] \leq \Pr \left[\sum_{i \in [t]} Y_i \geq \frac{1}{Mr} \right] \leq \exp(-\log(t/\delta)) = \delta/t.$$

A union bound completes the proof. ■

Lemma 2 *If t , T , and M satisfy the conditions of Lemma 1, the event \mathcal{E}_h holds, and vector y of at most unit norm formed by taking a vector $y' \in C(A)$ of at most unit norm and replacing y'_i with 0 whenever $u_i \geq T$, then for any $\ell \geq 2$, let $\mathcal{E}_L(\epsilon)$ be the event that $|\|\Phi D y\|_2^2 - \|y\|^2| \leq \epsilon$. Then*

$$\Pr[\mathcal{E}_L(\epsilon)] \geq 1 - \epsilon^{-\ell} Z^\ell,$$

where $Z = \max\{\sqrt{\ell} \cdot \|y\|_2 \sqrt{2/(Mr)}, \ell \cdot 2/(Mr)\}$. Here y can depend on Φ , but not on D .

Proof: We will use the following theorem, due to Hanson and Wright.

Theorem 3 [27] *Let $z \in \mathbb{R}^n$ be a vector of i.i.d. ± 1 random values. For any symmetric $B \in \mathbb{R}^{n \times n}$ and $\ell \geq 2$,*

$$\mathbf{E} [|z^\top B z - \text{tr}(B)|^\ell] \leq (CZ)^\ell, \text{ where } Z \equiv \max\{\sqrt{\ell} \|B\|_F, \ell \cdot \|B\|_2\},$$

and $C > 0$ is a universal constant.

We will use Theorem 3 to prove a bound on the ℓ 'th moment of $\|\Phi D y\|_2^2$ for large ℓ . Note that $\|\Phi D y\|^2$ can be written as $z^\top B z$, where z has entries from the diagonal of D , and $B \in \mathbb{R}^{n \times n}$ has $B_{ii'} = y_i y_{i'} I_{h(i)=h(i')}$. Here $\text{tr}(B) = \|y\|^2$.

Our analysis uses some ideas from the proofs for Lemmas 7 and 8 of [29].

Since by assumption event \mathcal{E}_h of Lemma 1 occurs, and for a vector $y \in C(A)$ of at most unit norm $y_{i'}^2 \leq u_{i'}$ for all i' , we have for $j \in [t]$ that $\sum_{i' \in h^{-1}(j)} y_{i'}^2 \leq 2/(Mr)$. Hence

$$\|B\|_F^2 = \sum_{i, i'} (y_i y_{i'})^2 I_{h(i)=h(i')} = \sum_{i \in [n]} y_i^2 \sum_{i' \in h^{-1}(h(i))} y_{i'}^2 \leq \sum_{i \in [n]} y_i^2 2/(Mr) \leq \|y\|_2^2 \cdot 2/(Mr). \quad (3)$$

For $\|B\|_2$, observe that for given $j \in [t]$, $y^{(j)} \in \mathbb{R}^n$ with $y_i^{(j)} = y_i I_{h(i)=j}$ is an eigenvector of B with eigenvalue $\|y^{(j)}\|^2$, and the set of such eigenvectors spans the column space of B . It follows that

$$\|B\|_2 = \max_j \|y^{(j)}\|^2 = \sum_{i' \in h^{-1}(j)} y_{i'}^2 \leq 2/(Mr).$$

Putting this and (3) into the Z of Theorem 3, we have,

$$Z \leq \max\{\sqrt{\ell}\|B\|_F, \ell \cdot \|B\|_2\} \leq \max\{\sqrt{\ell} \cdot \|y\|_2 \sqrt{2/(Mr)}, \ell \cdot 2/(Mr)\}$$

By a Markov bound applied to $|z^\top Bz - \text{tr}(B)|^\ell$,

$$\Pr[|\|\Phi D y\|_2^2 - \|y\|^2| \geq \epsilon] \leq \epsilon^{-\ell} Z^\ell.$$

■

3.2 Handling vectors with large entries

Now consider a unit vector y with $y_i = 0$ if $u_i \leq T$, for all $i \in [n]$.

Lemma 4 *For a parameter $T > 0$, let $H \subset [n]$ denote the subset of rows i for which $u_i \geq T$. Let \mathcal{E}_B denote the event that for all $i, j \in H$ with $i \neq j$, $h(i) \neq h(j)$. Then for $T \geq 4r/\sqrt{t}$, $\Pr[\mathcal{E}_B] \geq 15/16$.*

Proof: Since the columns of U are orthonormal, $\sum_{i \in [n]} u_i = \sum_{i \in [n]} \|U_{i,*}\|^2 = r$. Therefore $|H| \leq r/T \leq \sqrt{t}/4$. For fixed $i \neq j$, $\Pr[h(i) = h(j)] = 1/t$. Therefore by a union bound, the probability that some $i \neq j$ have $h(i) = h(j)$ is at most $|H|^2/t \leq 1/16$. Thus $\Pr[\mathcal{E}_B] \geq 15/16$. ■

Given event \mathcal{E}_B , we have that for any y of the form above,

$$\|y\|_2 = \|\Phi D y\|_2.$$

More generically, let \mathcal{E}_H be the event that for such a y , $\|\Phi D y\|^2 \leq (1 + \epsilon)\|y\|^2$.

3.3 Handling all vectors

We have seen that ΦD preserves the norms for vectors with small entries (Lemma 2) and large entries (Lemma 4). Before proving a general bound, we need to prove a bound on the “cross terms”.

First, a bound on a key contributor to the cross terms.

Lemma 5 *For fixed unit $y \in \mathbb{R}^n$, let vector y^L have $y_i^L = y_i I_{u_i \leq T}$ and vector $y^H \equiv y - y^L$. Let H be the set of indices of nonzero coordinates of y^H , and $|H| \leq s \equiv r/T$. Let $G \equiv \{i \mid i \in h^{-1}(i'), i' \in H\}$. Let $y^{L'} \in \mathbb{R}^n$ have $y_i^{L'} = y_i^L I_{i \in G}$. For $\delta' > 0$, suppose $T \leq \epsilon^2 / \log(1/\delta')$, and $t \geq 6r \log(1/\delta')/\epsilon^4$. Then it holds with failure probability at most δ' that $\|y^{L'}\|^2 \leq 2\epsilon^2$.*

Proof: To bound $\|y^{L'}\|$, let X_i denote the random variable $y_i^2 I_{i \in G}$. We will apply Bernstein’s inequality to $\sum_i Y_i$ where $Y_i \equiv X_i - \mathbf{E}[X_i]$. We have $|Y_i| \leq T$, and with that bound, Bernstein’s inequality states that

$$\log \Pr \left[\sum_{i \in [t]} Y_i \geq z \right] \leq \frac{-z^2/2}{\sum_i \mathbf{E}[Y_i^2] + zT/3}. \quad (4)$$

We have $\mathbf{E}[Y_i^2] \leq \mathbf{E}[X_i^2] \leq y_i^2 T s/t$ so that $\sum_i \mathbf{E}[Y_i^2] \leq T s/t = r/t$. Hence, (4) becomes

$$\log \Pr \left[\sum_{i \in [t]} Y_i \geq z \right] \leq \frac{-z^2/2}{r/t + zT/3}.$$

Using that $T \leq \epsilon^2 / \log(1/\delta')$ and that $t \geq 6r \log(1/\delta') / \epsilon^4$,

$$\log \Pr \left[\sum_{i \in [t]} Y_i \geq \epsilon^2 \right] \leq \frac{-\epsilon^4}{\epsilon^4 / \log(1/\delta')} = -\log(1/\delta').$$

Now since $\sum_i \mathbf{E}[X_i] \leq s/t \leq \epsilon^2$,

$$\Pr \left[\sum_{i \in [t]} X_i \geq 2\epsilon^2 \right] \leq \Pr \left[\sum_{i \in [t]} Y_i \geq \epsilon^2 \right] \leq \exp(-\log(1/\delta')) = \delta'.$$

Hence

$$\Pr[\|y^{L'}\|^2 \leq 2\epsilon^2] \geq 1 - \delta',$$

as claimed. ■

Lemma 6 *For fixed unit $y \in C(A)$, and using the notation of the previous lemma, suppose the event of the previous lemma holds, and also event \mathcal{E}_H , and also $\mathcal{E}_L(\epsilon^2)$ holds for $y^{L'}$. Then $|(y^H)^\top D\Phi^\top \Phi D y^L| \leq \sqrt{3}\epsilon(1 + \epsilon)$.*

Proof: From the assumptions of \mathcal{E}_H and $\mathcal{E}_L(\epsilon^2)$ for $y^{L'}$,

$$(y^H)^\top D\Phi^\top \Phi D y^L = (y^H)^\top D\Phi^\top \Phi D y^{L'} \leq \|\Phi D y^H\| \|\Phi D y^{L'}\| \leq (1 + \epsilon) \sqrt{\|y^{L'}\|^2 + \epsilon^2}.$$

Using the bound on $\|y^{L'}\|$ from the previous lemma, assumed to hold, the lemma follows. ■

Lemma 7 *Given any $K \in (0, 1)$, for fixed $y \in C(A)$, assuming the parameter conditions for Lemmas 1, 2, 4, and 6 hold, and events \mathcal{E}_H and \mathcal{E}_h , then for embedding dimension $t = O((r/\epsilon)^4 \log^2(r/\epsilon))$, with probability at least $1 - K^r$, $\|\Phi D y\|_2 = (1 \pm \epsilon)\|y\|_2$.*

Proof: Write $y = y^H + y^L$, where y^L has $y_i^L = y_i I_{u_i \leq T}$. We bound

$$\|\Phi D y\|_2^2 = \|\Phi D y^H\|_2^2 + \|\Phi D y^L\|_2^2 + 2y^H D\Phi^\top \Phi D y^L. \quad (5)$$

We apply Lemma 2 with $\ell = r$. To satisfy the other conditions of the lemmas, we need to choose T , t , δ' , and M such that

1. For given $C > 0$, $(2/(M\epsilon^2))^{r/2} \leq C^r$. So $M = \Omega(\epsilon^{-2})$ suffices for this;
2. $T \leq 1/(6Mr \log(t/\delta)) = O(1/(\epsilon^2 r \log(t)))$ for fixed δ ;
3. $T \geq 4r/\sqrt{t}$;
4. $\delta' \leq C_\delta^r$, for small enough $C_\delta > 0$;
5. $T \leq \epsilon^2 / \log(t/\delta') = O(\epsilon^2 / (r \log(t)))$;
6. $t \geq 6r \log(1/\delta') / \epsilon^4 = \Omega(r^2 / \epsilon^4)$;

There are values $T = O(\epsilon^2 / (r \log(r/\epsilon)))$ and $t = O((r/\epsilon)^4 \log^2(r/\epsilon))$ that satisfy these conditions.

With the first two conditions and the event \mathcal{E}_h , Lemma 2 implies that $|\|\Phi D y^L\|^2 - \|y^L\|^2| \leq \epsilon$ with failure probability C^{-r} .

The third condition is needed to apply Lemma 4, which implies that \mathcal{E}_B , and hence \mathcal{E}_H , hold with constant probability.

The remaining conditions imply by Lemma 5 that $\|y^{L'}\|^2 \leq 2\epsilon^2$ with failure probability at most δ' . This and the first two conditions (with insignificantly larger M), and \mathcal{E}_h , imply from Lemma 2 that $\mathcal{E}_L(\epsilon^2)$

holds for $y^{L'}$ with failure probability at most C^r . (Here we can consider the hash function in Lemma 2 to be restricted to the buckets containing the indices in H , but the results of the lemma still apply.) With Lemma 6, this implies that $|(y^H)^\top D \Phi^\top \Phi D y^L| \leq \sqrt{3}\epsilon(1 + \epsilon)$.

That is, with failure probability at most $2C^r + C_\delta^r \leq K^r$, where C and C_δ are chosen less than $K/3$, we have $|\|\Phi D y\|_2^2 - \|y\|_2^2| \leq 2\epsilon + \sqrt{3}\epsilon(1 + \epsilon)$. Adjusting ϵ gives the result. \blacksquare

Lemma 8 *Suppose L is an r -dimensional subspace of \mathbb{R}^n , and $B : \mathbb{R}^n \rightarrow \mathbb{R}^k$ is a linear map. If for any fixed $x \in L$, $\|Bx\|_2^2 = (1 \pm \epsilon/6)\|x\|_2^2$ with probability at least $1 - p$, then there is a constant $C_{sub} > 0$ for which with probability at least $1 - p(C_{sub})^r$, for all $x \in L$, $\|Bx\|_2^2 = (1 \pm \epsilon)\|x\|_2^2$.*

Proof: We will need the following standard lemmas for making a net argument. Let S^{r-1} be the unit sphere in \mathbb{R}^r and let E be the set of points in S^{r-1} defined by

$$E = \left\{ w : w \in \frac{\gamma}{\sqrt{r}} \mathbb{Z}^r, \|w\|_2 \leq 1 \right\},$$

where \mathbb{Z}^r is the r -dimensional integer lattice.

Fact 9 (Lemma 4 of [4]) $|E| \leq e^{cr}$ for $c = (\frac{1}{\gamma} + 2)$.

Fact 10 (Lemma 4 of [4]) *For any $r \times r$ matrix J , if for every $u, v \in E$ we have $|u^\top J v| \leq \epsilon$, then for every unit vector w , we have $|w^\top J w| \leq \frac{\epsilon}{(1-\gamma)^2}$.*

Let $U \in \mathbb{R}^{n \times r}$ be such that the columns are orthonormal and the column space equals L . Let I_r be the $r \times r$ identity matrix. Define $J = U^\top B^\top B U - I_r$. Consider the set E in Fact 9 and Fact 10. Then, for any $x, y \in E$, we have by the statement of the lemma that with probability at least $1 - 3p$, $\|B U x\|_2^2 = (1 \pm \epsilon/6)\|U x\|_2^2$, $\|B U y\|_2^2 = (1 \pm \epsilon/6)\|U y\|_2^2$, and $\|B U(x+y)\|_2^2 = (1 \pm \epsilon/6)\|U(x+y)\|_2^2 = (1 \pm \epsilon/6)(\|U x\|_2^2 + \|U y\|_2^2 + 2\langle U x, U y \rangle)$. Since $\|U x\|_2 \leq 1$ and $\|U y\|_2 \leq 1$, it follows that $|x J y| \leq \epsilon/2$. By Fact 9, for $\gamma = 1 - 1/\sqrt{2}$ and sufficiently large C_{sub} , we have by a union bound that with probability at least $1 - p C_{sub}^r$ for a constant $C_{sub} > 0$, that $|x J y| \leq \epsilon/2$ for every $x, y \in E$. Hence, with this probability, by Fact 10, $|w^\top J w| \leq \epsilon$ for every unit vector w , which by definition of J means that for all $y \in L$, $\|B y\|_2^2 = (1 \pm \epsilon)\|y\|_2^2$. \blacksquare

The following is our main theorem in this section.

Theorem 11 *There is $t = O((r/\epsilon)^4 \log^2(r/\epsilon))$ such that with probability at least $9/10$, ΦD is a subspace embedding matrix for A ; that is, for all $y \in C(A)$, $\|\Phi D y\|_2 = (1 \pm \epsilon)\|y\|_2$. The embedding ΦD can be applied in $O(\text{nnz}(A))$ time.*

Proof: We set $\delta = 1/60$ in Lemma 1, so that \mathcal{E}_h occurs with probability at least $1 - 1/60$. Next, the event \mathcal{E}_B of Lemma 4 occurs with probability at least $15/16$. By a union bound, both events \mathcal{E}_H and \mathcal{E}_h occur with probability at least $1 - 1/60 - 1/16$. Conditioned on this, by Lemma 7, for any fixed $y \in C(A)$, with probability at least $1 - K^r$, $\|\Phi D y\|_2 = (1 \pm \epsilon)\|y\|_2$, where $K > 0$ can be chosen arbitrarily small. Hence, by Lemma 8, and choosing $K < 1/60 C_{sub}$, with probability at least $1 - 60^{-r}$, $\|\Phi D y\|_2 = (1 \pm 6\epsilon)\|y\|_2$ for all $y \in C(A)$. Hence, by a union bound, with probability at least $9/10$, $\|\Phi D y\|_2 = (1 \pm 6\epsilon)\|y\|_2$ for all $y \in L$. The theorem follows by rescaling ϵ by a constant factor. \blacksquare

4 Generalized Sparse Embedding Matrices

4.1 Johnson-Lindenstrauss transforms

We start with a theorem of Kane and Nelson [29], restated here in our notation. We also present a simple corollary that we need concerning very low dimensional subspaces. Let $\epsilon > 0$, $k = \Theta(\epsilon^{-2} \log(r/\epsilon))$, and $v = \Theta(\epsilon^{-1})$ be such that v divides k . Let $B : \mathbb{R}^n \rightarrow \mathbb{R}^k$ be defined as follows. We view B as the concatenation (meaning, we stack the rows on top of each other) of matrices $\sqrt{v/k} \cdot \Phi_1 \cdot D_1, \dots, \sqrt{v/k} \cdot \Phi_{k/v} \cdot D_{k/v}$, each $\Phi_i \cdot D_i$ being a linear map from \mathbb{R}^n to \mathbb{R}^v , which is an independently chosen sparse embedding matrix of Section 3 with associated hash function $h_i : [n] \rightarrow [v]$.

Theorem 12 ([29]) For any $\varepsilon > 0$ and constant $C_{KN} > 0$, there are $k = \Theta(\varepsilon^{-2} \log(r/\varepsilon))$ and $v = \Theta(\varepsilon^{-1})$ with $v \mid k$ for which for any fixed $x \in \mathbb{R}^n$, a randomly chosen B of the form above satisfies $\|Bx\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$ with probability at least $1 - (\varepsilon/r)^{C_{KN}}$.

Corollary 13 Suppose L is an $O(\log(r/\varepsilon))$ -dimensional subspace of \mathbb{R}^n . Let $C_{subKN} > 0$ be any constant. Then for any $0 < \varepsilon < 1$, there are $k = \Theta(\varepsilon^{-2} \log(r/\varepsilon))$ and $v = \Theta(\varepsilon^{-1})$ with $v \mid k$ for which with probability at least $1 - (\varepsilon/r)^{C_{subKN}}$, for all $y \in L$, $\|By\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$.

Proof: We use Theorem 12 together with Lemma 8; for the latter, we need that for any fixed $y \in L$, $\|By\|_2^2 = (1 \pm \varepsilon/6)\|y\|_2^2$ with probability at least $1 - p$. By Theorem 12, we have this for $p = (\varepsilon/r)^{C_{KN}}$ for an arbitrarily large constant $C_{KN} > 0$. Hence, by Lemma 8, there is a constant $C_{sub} > 0$ so that with probability at least $1 - (C_{sub})^{O(\log(r/\varepsilon))}(\varepsilon/r)^{C_{KN}} = 1 - (\varepsilon/r)^{C_{subKN}}$, for all $y \in L$, $\|By\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$. Here we use that $C_{KN} > 0$ can be made arbitrarily large, independent of C_{sub} . ■

4.2 The construction

We now define a *generalized sparse embedding matrix* S . Let $A \in \mathbb{R}^{n \times d}$ with rank r .

Let $k = \Theta(\varepsilon^{-2} \log(r/\varepsilon))$ and $v = \Theta(\varepsilon^{-1})$, with $v \mid k$, be such that Theorem 12 and Corollary 13 apply with parameters k and v , for a sufficiently large constant $C_{subKN} > 0$. Further, let

$$t = C_t r \varepsilon^{-4} \log(r/\varepsilon)(r + \log(1/\varepsilon)),$$

where $C_t > 0$ is a sufficiently large absolute constant, and suppose that $k \mid t$. Let $q = t/k$.

Let $h : [n] \rightarrow [q]$ be a random hash function. For $i = 1, 2, \dots, q$, define $a_i = |h^{-1}(i)|$. Note that $\sum_{i=1}^q a_i = n$.

We choose independent matrices $B^{(1)}, \dots, B^{(q)}$, with each $B^{(i)}$ as in Theorem 12 with parameters k and v . Here $B^{(i)}$ is a $k \times a_i$ matrix. Finally, let P be an $n \times n$ permutation matrix which, when applied to a matrix A , maps the rows of A in the set $h^{-1}(1)$ to the set of rows $\{1, 2, \dots, a_1\}$, maps the rows of A in the set $h^{-1}(2)$ to the set of rows $\{a_1 + 1, \dots, a_1 + a_2\}$, and for a general $i \in [q]$, maps the set of rows of A in the set $h^{-1}(i)$ to the set of rows $\{a_1 + a_2 + \dots + a_{i-1} + 1, \dots, a_1 + a_2 + \dots + a_i\}$.

The map S is defined to be the product of a block-diagonal matrix and the matrix P :

$$S \equiv \begin{bmatrix} B^{(1)} & & & \\ & B^{(2)} & & \\ & & \ddots & \\ & & & B^{(q)} \end{bmatrix} \cdot P$$

Lemma 14 $S \cdot A$ can be computed in $O(\text{nnz}(A)(\log(r/\varepsilon))/\varepsilon)$ time.

Proof: As P is a permutation matrix, $P \cdot A$ can be computed in $O(\text{nnz}(A))$ time and has the same number of non-zero entries of A . For each non-zero entry of $P \cdot A$, we multiply it by $B^{(i)}$ for some i , which takes $O(k/v) = O(\log(r/\varepsilon)/\varepsilon)$ time. Hence, the total time to compute $S \cdot A$ is $O(\text{nnz}(A)(\log(r/\varepsilon))/\varepsilon)$. ■

4.3 Analysis

We adapt the analysis given for sparse embedding matrices to generalized sparse embedding matrices. Again let $U \in \mathbb{R}^{n \times r}$ have columns that form an orthonormal basis for the column space $C(A)$. Let $U_{1,*}, \dots, U_{n,*}$ be the rows of U , and let $u_i \equiv \|U_{i,*}\|^2$. We set the parameter:

$$T = \frac{\varepsilon^2}{12C_T \log(10k/v)(r + \log(1/\varepsilon))},$$

where C_T is a sufficiently large absolute constant.

4.3.1 Vectors with small entries

Consider a fixed vector y of at most unit norm formed by taking a vector $y' \in C(A)$ of at most unit norm, and replacing y'_i with 0 whenever $u_i \geq T$. Since $y_i^2 \leq u_i$, this implies that $\|y\|_\infty^2 \leq T$. Since P is a permutation matrix, we have $\|Py\|_\infty^2 \leq T$.

In this case, we can reduce the analysis to that of a sparse embedding matrix. Indeed, observe that the matrix $B^{(i)}$ is the concatenation of matrices $\Phi_1^{(i)} D_1^{(i)}, \dots, \Phi_{k/v}^{(i)} D_{k/v}^{(i)}$, where each $\Phi_j^{(i)} D_j^{(i)}$ is a sparse embedding matrix. Now fix a value $j \in [k/v]$ and consider the block-diagonal matrix N_j :

$$N_j \equiv \begin{bmatrix} \Phi_j^{(1)} D_j^{(1)} & & & \\ & \Phi_j^{(2)} D_j^{(2)} & & \\ & & \ddots & \\ & & & \Phi_j^{(q)} D_j^{(q)} \end{bmatrix} \cdot P$$

Lemma 15 N_j is a random sparse embedding matrix with $qv = tv/k$ rows and n columns.

Proof: N_j has a single non-zero entry in each column, and the value of this non-zero entry is random in $\{+1, -1\}$. Hence, it remains to show that the distribution of locations of the non-zero entries of N_j is the same as that in a sparse embedding matrix. This follows from the distribution of the values a_1, \dots, a_q . ■

For $j = 1, \dots, k/v$, let \mathcal{E}_h^j be the event \mathcal{E}_h of Lemma 1, applied to matrix N_j .

We set δ in Lemma 1 to be $v/(10k)$ to conclude that by a union bound, $\cap_{j=1}^{k/v} \mathcal{E}_h^j$ occurs with probability at least $9/10$.

We apply Lemma 2 on $N_j y$ for each $j = 1, \dots, k/v$. We will set the parameter “ ℓ ” there to be $r + \log 1/\varepsilon$. The parameter “ M ” in Lemma 2 is set to be $2C_T(1/\varepsilon^2 + \log(1/\varepsilon)/(r\varepsilon^2))$, and the parameter “ t ” in Lemma 2 is equal to $qv = \Omega(r^2\varepsilon^{-3} + r\varepsilon^{-3} \log(1/\varepsilon))$, which in turn is at least $Mr^2 = \Theta(r^2/\varepsilon^2 + r \log(1/\varepsilon)/\varepsilon^2)$ by our choice of q and v . Notice that the number of rows of N_j is indeed qv by Lemma 15. Also, the parameter “ δ ” in Lemma 1 equals $v/(10k)$. We set the parameter “ T ” in Lemma 2 to be

$$T = \frac{1}{6Mr \log 1/\delta} = \frac{\varepsilon^2}{12C_T \log(10k/v)(r + \log(1/\varepsilon))},$$

which agrees with our aforementioned value of T , and satisfies the requirement of Lemma 2 (and also of Lemma 1). Therefore, we can indeed apply Lemma 2.

By Lemma 2, we have for each $j \in [k/v]$:

$$\Pr[| \|N_j y\|_2^2 - \|y\|_2^2 | \geq \varepsilon] \leq \left(\frac{2\ell}{Mr\varepsilon^2} \right)^{\ell/2} = \left(\frac{2(r + \log 1/\varepsilon)}{Mr\varepsilon^2} \right)^{(r + \log 1/\varepsilon)/2} = \left(\frac{2}{M\varepsilon^2} + \frac{2 \log 1/\varepsilon}{Mr\varepsilon^2} \right)^{(r + \log 1/\varepsilon)/2}.$$

Plugging in our choice of M , for $C_T > 0$ a sufficiently large constant, this is

$$\Pr[| \|N_j y\|_2^2 - \|y\|_2^2 | \geq \varepsilon] \leq \left(\frac{2}{C_T} \right)^{(r + \log 1/\varepsilon)/2} = O(\varepsilon^2) \left(\frac{2}{C_T} \right)^{r/2},$$

and so by a union bound,

$$\Pr[\exists j, | \|N_j y\|_2^2 - \|y\|_2^2 | \geq \varepsilon] \leq (k/v) \cdot O(\varepsilon^2) \left(\frac{2}{C_T} \right)^{r/2} = O(\varepsilon^{-1} \log(r/\varepsilon)) \cdot O(\varepsilon^2) \cdot \left(\frac{C_T}{2} \right)^{-r/2} = O(1) \cdot \left(\frac{C_T}{3} \right)^{-r/2}.$$

Since

$$\|Sy\|_2^2 = \frac{v}{k} \sum_{j=1}^{k/v} \|N_j y\|_2^2,$$

it follows that

$$\Pr[| \|Sy\|_2^2 - \|y\|_2^2 | \geq \varepsilon] = O(1) \cdot (C_T/3)^{-r/2}.$$

4.3.2 Vectors with large entries

Consider the set H of i for which $u_i > T$. Then by our choice of T ,

$$|H| \leq \frac{r}{T} = \frac{12C_T r \log(10k/v)(r + \log(1/\varepsilon))}{\varepsilon^2} = O(r\varepsilon^{-2}(\log(1/\varepsilon) + \log \log r)(r + \log(1/\varepsilon))).$$

The following is a standard non-weighted balls-and-bins analysis.

Lemma 16 *For a large enough constant $C_t > 0$, with probability at least $1 - 1/r$, for all $i \in [q]$, $|h^{-1}(i) \cap H| \leq C_t \log(r/\varepsilon)$.*

Proof: For any given $i \in [q]$, $\mathbf{E}[|h^{-1}(i) \cap H|] = |H|/q = |H|k/t$. Using that

$$t/k = \Omega(r\varepsilon^{-2}(r + \log(1/\varepsilon))),$$

we have that this expectation is $O(\log(1/\varepsilon) + \log \log r) = O(\log(r/\varepsilon))$. Hence, by a Chernoff bound, for a constant $C_t > 0$,

$$\Pr[|h^{-1}(i) \cap H| > C_t \log(r/\varepsilon)] \leq e^{-\Theta(\log(r/\varepsilon))} = \frac{1}{rq},$$

The lemma now follows by a union bound over all $i \in [q]$. ■

Let \mathcal{E}_{nw} be the event of Lemma 16, and assume \mathcal{E}_{nw} occurs, which happens with probability at least $1 - 1/r$.

For $i = 1, 2, \dots, q$, let L^i be the at most $C_t \log(r/\varepsilon)$ -dimensional subspace which is the restriction of the column space $C(A)$ to coordinates j with $h^{-1}(j) = i$ and $u_j \geq T$. By Corollary 13, for any fixed i , with probability at least $1 - (\varepsilon/r)^{C_{subKN}}$, for all $y \in L^i$, $\|Sy\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$. By a union bound and sufficiently large $C_{subKN} > 0$, this holds for all $i \in [q]$ with probability at least $1 - q(\varepsilon/r)^{C_{subKN}} > 1 - 1/r$. Let \mathcal{E}_s be the event that this occurs for all i .

Then, conditioned on \mathcal{E}_{nw} and \mathcal{E}_s , which jointly occur with probability at least $1 - 2/r$, it follows that $\|Sy\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$ for all vectors y obtained by taking a vector $y' \in C(A)$ and replacing the coordinates y'_j with 0 for those j for which $u_j < T$.

4.4 Putting it all together

Now consider any unit vector y in $C(A)$, and write it as $y^H + y^L$, where $y_i^H = 0$ for coordinates i for which $u_i < T$, while $y_i^L = 0$ for coordinates i for which $u_i \geq T$. We seek to bound $\langle Sy^H, Sy^L \rangle$. For notational

convenience, define the block-diagonal matrix \tilde{N}_j to be the matrix

$$\tilde{N}_j \equiv \begin{bmatrix} 0 & & & & & & & & & & \\ \dots & & & & & & & & & & \\ 0 & & & & & & & & & & \\ \Phi_j^{(1)} D_j^{(1)} & & & & & & & & & & \\ 0 & & & & & & & & & & \\ \dots & & & & & & & & & & \\ 0 & & & & & & & & & & \\ & 0 & & & & & & & & & \\ & \dots & & & & & & & & & \\ & 0 & & & & & & & & & \\ & \Phi_j^{(2)} D_j^{(2)} & & & & & & & & & \\ & 0 & & & & & & & & & \\ & \dots & & & & & & & & & \\ & 0 & & & & & & & & & \\ & & \ddots & & & & & & & & \\ & & & 0 & & & & & & & \\ & & & \dots & & & & & & & \\ & & & 0 & & & & & & & \\ & & & & \Phi_j^{(q)} D_j^{(q)} & & & & & & \\ & & & & 0 & & & & & & \\ & & & & \dots & & & & & & \\ & & & & 0 & & & & & & \end{bmatrix} \cdot P$$

Then $S = \sqrt{v/k} \cdot \sum_{j=1}^{k/v} \tilde{N}_j$. Notice that since the set of non-zero rows of \tilde{N}_j and $\tilde{N}_{j'}$ are disjoint for $j \neq j'$,

$$\langle Sy^H, Sy^L \rangle = \frac{v}{k} \sum_{j=1}^{k/v} \langle \tilde{N}_j y^H, \tilde{N}_j y^L \rangle = \frac{v}{k} \sum_{j=1}^{k/v} \langle N_j y^H, N_j y^L \rangle, \quad (6)$$

where by Lemma 15, each N_j is a sparse embedding matrix with $qv = t \cdot v/k$ rows and n columns.

We will apply Lemma 6 to bound each summand $\langle N_j y^H, N_j y^L \rangle$. We condition on events \mathcal{E}_{nw} , \mathcal{E}_s , and $\cap_{j=1}^{k/v} \mathcal{E}_h^j$ jointly occurring, which by a union bound happens with probability at least $9/10 - 2/r$.

Also, let $s = r/T$. We apply Lemma 6 with our values of s, t, T , and M specified above, together with the value $\delta' = \varepsilon C_T^{-r}$. To apply Lemma 5 and Lemma 6, we need

1. t, T , and M to satisfy Lemma 1;

$$2. T \leq \frac{\varepsilon^2}{\log(t/\delta')} = \frac{\varepsilon^2}{C_T r + \log(t/\varepsilon)};$$

$$3. t \geq \frac{6r \log(t/\delta')}{\varepsilon^4} = \frac{6r(C_T r + \log(t/\varepsilon))}{\varepsilon^4}.$$

We have already shown that the first condition holds. Setting $t = C_t r^2 \varepsilon^{-4} \log(r/\varepsilon) + C_t r \varepsilon^{-4} \log(r/\varepsilon) \log(1/\varepsilon)$ for a large enough constant C_t , we have $\log(t/\varepsilon) \leq 6 \log r + 6 \log(1/\varepsilon)$ for r and $1/\varepsilon$ larger than a large enough constant. Plugging this choice of t into the above, we see that our choices of t and $T = \frac{\varepsilon^2}{12 C_T \log(10k/v)(r + \log(1/\varepsilon))}$ satisfy the remaining two conditions for r and $1/\varepsilon$ larger than a large enough constant.

Notice that \mathcal{E}_{nw} and \mathcal{E}_s occurring implies that the event \mathcal{E}_H of Lemma 6 occurs. Also, $\cap_{j=1}^{k/v} \mathcal{E}_h^j$ occurring implies that the event \mathcal{E}_h of Lemma 1 occurs. Therefore, applying Lemmas 5 and 6, given the occurrence of these events we have that with probability at least $1 - \delta'$,

$$|\langle N_j y^H, N_j y^L \rangle| \leq \sqrt{3} \varepsilon (1 + \varepsilon).$$

By a union bound over $j \in [k/v]$, we have that this occurs with probability at least $1 - (k/v)\delta'$ for all $j \in [k/v]$. Therefore, with probability at least $1 - (k/v)\delta' = 1 - O(1) \cdot C_T^{-r}$, it holds that $|\langle Sy^H, Sy^L \rangle| = O(\varepsilon)$. In this case, with probability at least $1 - O(1) \cdot (C_T/3)^{-r/2} - O(1) \cdot C_T^{-r}$,

$$\begin{aligned} \|Sy\|_2^2 &= \|Sy^H\|_2^2 + \|Sy^L\|_2^2 + 2\langle Sy^H, Sy^L \rangle \\ &= (1 \pm O(\varepsilon))\|y^H\|_2^2 + (1 \pm O(\varepsilon))\|y^L\|_2^2 \pm O(\varepsilon) \\ &= (1 \pm O(\varepsilon))\|y\|_2^2, \end{aligned}$$

where the last equality uses that y is a unit vector.

The following is our main theorem in this section.

Theorem 17 *With probability at least $9/10$, for $t = O(r\varepsilon^{-4} \log(r/\varepsilon)(r + \log(1/\varepsilon)))$, S is an embedding matrix for A ; that is, for all $y \in C(A)$, $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$. S can be applied to A in $O(\text{nnz}(A)(\log r)/\varepsilon)$ time.*

Proof: Events $\mathcal{E}_{nw}, \mathcal{E}_s$ and $\cap_{j=1}^{k/v} \mathcal{E}_h^j$ jointly occur with probability at least $9/10 - 2/r$. Conditioned on their joint occurrence, the above analysis shows that for any fixed $y \in C(A)$, $\|Sy\|_2 = (1 \pm O(\varepsilon))\|y\|_2$ with probability at least $1 - O(1) \cdot (C_2/3)^{-r/2}$, where $C_2 > 0$ is an arbitrarily large constant. Applying Lemma 8, with probability at least $1 - O(1) \cdot (C_2/3)^{-r/2} \cdot C_{sub}^r > 1 - 1/r$, for all $y \in C(A)$, $\|Sy\|_2 = (1 \pm O(\varepsilon))\|y\|_2$. Hence, by a union bound, with probability at least $4/5$, for all $y \in C(A)$, $\|Sy\|_2 = (1 \pm O(\varepsilon))\|y\|_2$. The theorem follows by rescaling ε by a constant factor. \blacksquare

5 Approximating the Leverage Scores

Let $A \in \mathbb{R}^{n \times d}$ with rank r . Let $U \in \mathbb{R}^{n \times r}$ be an orthonormal basis for $C(A)$. In [18] it was shown how to obtain a $(1 \pm \varepsilon)$ -approximation u'_i to u_i for all $i \in [n]$, for a constant $\varepsilon > 0$, in time $O(nd \log n) + O(d^3 \log n \log d)$. Here we improve the running time of this task as follows. We state the running time for constant ε , though for general ε the running time would be $O(\text{nnz}(A) \log n) + \text{poly}(r\varepsilon^{-1} \log n)$.

Theorem 18 *For any constant $\varepsilon > 0$, there is an algorithm which with probability at least $2/3$, outputs a vector (u'_1, \dots, u'_n) so that for all $i \in [n]$, $u'_i = (1 \pm \varepsilon)u_i$. The running time is*

$$O(\text{nnz}(A) \log n + r^3 \log^2 r + r^2 \log n).$$

The success probability can be amplified by independent repetition and taking the coordinate-wise median of the vectors u' across the repetitions.

Proof: We first run the algorithm of Theorem 2.6 and Theorem 2.7 of [7]. The first theorem gives an algorithm which outputs the rank r of A , while the second theorem gives an algorithm which also outputs the indices i_1, \dots, i_r of linearly independent columns of A . The algorithm takes $O(\text{nnz}(A) \log d) + O(r^3)$ time and succeeds with probability at least $1 - O(\log d)/d^{1/3}$. Hence, in what follows, we can assume that A has full rank.

We follow the same procedure as Algorithm 1 in [18], using our improved subspace embedding. The proof of [18] proceeds by choosing a subspace embedding Π_1 , computing $\Pi_1 A$, then computing a change of basis matrix R so that $\Pi_1 A R$ has orthonormal columns. The analysis there then shows that the row norms $\|(AR)_{i,*}\|_2^2$ are equal to $u_i(1 \pm \varepsilon)$. To obtain these row norms quickly, an $r \times O(\log n)$ Johnson-Lindenstrauss matrix Π_2 is sampled, and one first computes $R\Pi_2$, followed by $A(R\Pi_2)$. Using a fast Johnson-Lindenstrauss transform Π_1 , one can compute $\Pi_1 A$ in $O(nr \log n)$ time. Π_1 has $O(r \log n \log r)$ rows, and one can compute the $r \times r$ matrix R in $O(r^3 \log n \log r)$ time by computing a QR-factorization. Computing $R\Pi_2$ can be done in $O(r^2 \log n)$ time, and computing $A(R\Pi_2)$ can be done in $O(\text{nnz}(A) \log n)$ time.

Our only change to this procedure is to use a different matrix Π_1 , which is the composition of our subspace embedding matrix S of Theorem 17 with parameter $t = O(r^2 \log r)$, together with a fast Johnson-Lindenstrauss transform F . That is, we set $\Pi_1 = F \cdot S$. Here, F is an $O(r \log^2 r) \times t$ matrix, see Section

2.3 of [18] for an instantiation of F . Then, $S \cdot A$ can be computed in $O(\text{nnz}(A) \log r)$ time by Lemma 14. Moreover, $F \cdot (SA)$ can be computed in $O(t \cdot r \log r) = O(r^3 \log^2 r)$ time. One can then compute the matrix R above in $O(r^3 \log^2 r)$ time by computing a QR-factorization of FSA . Then one can compute $R\Pi_2$ in $O(r^2 \log n)$ time, and computing $A(R\Pi_2)$ can be done in $O(\text{nnz}(A) \log n)$ time. Hence, the total time is $O(\text{nnz}(A) \log n + r^3 \log^2 r + r^2 \log n)$ time.

Notice that by Theorem 17, with probability at least $4/5$, $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ for all $y \in C(A)$, and by Lemma 3 of [18], with probability at least $9/10$, $\|FSy\|_2 = (1 \pm \varepsilon)\|Sy\|_2$ for all $y \in C(A)$. Hence, $\|FSAx\|_2 = (1 \pm \varepsilon)^2\|Ax\|_2$ for all $x \in \mathbb{R}^d$ with probability at least $7/10$. There is also a small $1/n$ probability of failure that $\|(AR\Pi_2)_{i,*}\|_2 \neq (1 \pm \varepsilon)\|(AR)_{i,*}\|_2$ for some value of i . Hence, the overall success probability is at least $2/3$.

The rest of the correctness proof is identical to the analysis in [18]. ■

6 Least Squares Regression

Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be a matrix and vector for the regression problem: $\min_x \|Ax - b\|_2$. We assume $n > d$. Again, let r be the rank of A . We show that with probability at least $2/3$, we can find an x' for which

$$\|Ax' - b\|_2 \leq (1 + \varepsilon) \min_x \|Ax - b\|_2.$$

We will give two different algorithms, one using Theorem 11, and the other using Theorem 17.

Theorem 19 *The ℓ_2 -regression problem can be solved up to a $(1 + \varepsilon)$ -factor with probability at least $2/3$ in $O(\text{nnz}(A) + d^5 \varepsilon^{-4} \log^3(d/\varepsilon))$ time.*

Proof: By Theorem 11 applied to the column space of A adjoined with the vector b , it suffices to compute ΦDA and ΦDb and output $\text{argmin}_x \|\Phi DAx - \Phi Db\|_2$. We use the fact that $d \geq r$, and apply Theorem 11 with $t = O(d^4 \varepsilon^{-4} \log^2(d/\varepsilon))$.

The theorem implies that with probability at least $9/10$, all vectors y in the space spanned by the columns of A and b have their norms preserved up to a $(1 + \varepsilon)$ -factor. Notice that ΦDA and ΦDb can be computed in $O(\text{nnz}(A))$ time. Now we have a regression problem with $d' = O(d^4 \varepsilon^{-4} \log^2(d/\varepsilon))$ rows and d columns. Using the Fast Johnson-Lindenstrauss transform, this can be solved in $O(d'd \log(d/\varepsilon) + d^3 \varepsilon^{-1} \log d)$ time, see, Theorem 12 of [39]. The success probability is at least $9/10$. This is $O(d^5 \varepsilon^{-4} \log^3(d/\varepsilon))$ time. ■

Theorem 20 *The ℓ_2 -regression problem can be solved up to a $(1 + \varepsilon)$ -factor with probability at least $2/3$ in $O(\text{nnz}(A)(\log d + \log(1/\varepsilon)) + r^3 \varepsilon^{-2}(\log(r/\varepsilon))(\log(1/\varepsilon)))$ time.*

Proof: We first run the algorithm of Theorem 2.6 and Theorem 2.7 of [7]. The first theorem gives an algorithm which outputs the rank r of A , while the second theorem gives an algorithm which also outputs the indices i_1, \dots, i_r of linearly independent columns of A . The algorithm takes $O(\text{nnz}(A) \log d) + O(r^3)$ time and succeeds with probability at least $1 - O(\log d)/d^{1/3}$. Hence, in what follows, we can assume that A has full rank.

We then apply the subspace embedding of Theorem 11 to $A \circ b$, which takes $O(\text{nnz}(A) + r^4 \varepsilon^{-4} \log^2(r/\varepsilon))$ time and produces a matrix $A' = \Phi DA$ and vector $b' = \Phi Db$, where A' is $r' \times r$ and b' is $r' \times 1$, $r' = O(r^4 \varepsilon^{-4} \log^2(r/\varepsilon))$. Notice that $\text{nnz}(A') \leq \text{nnz}(A)$ by definition of ΦD . Observe that by Theorem 11, with probability at least $9/10$, for all $x \in \mathbb{R}^r$, $\|A'x - b'\|_2 = (1 \pm \varepsilon)\|Ax - b\|_2$.

Next, we need the following theorem of Dasgupta et al.

Fact 21 *(Theorem 5 of [9], restated in our notation) Let A be an $n \times d$ matrix of rank r . Let U be an orthonormal basis for $C(A)$, and let us randomly sample rows of A according to any distribution (p_1, \dots, p_n) for which for all i , $p_i \geq \min(1, q \cdot u_i/r)$, where $q \geq 32^2 r(r \ln(12/\varepsilon) + \ln(2/\delta))/(4\varepsilon^2)$. Then with probability $1 - \delta$, the following holds for all $x \in \mathbb{R}^d$:*

$$||SAx||_2 - \|Ax\|_2 \leq \varepsilon \|Ax\|_2,$$

where S is a diagonal matrix where $S_{i,i} = 1/\sqrt{p_i}$ if we sample row i , and $S_{i,i} = 0$ otherwise.

We run the algorithm of Theorem 18 on matrix A' to obtain a vector u' with $u'_i \geq \frac{1}{2}u_i$ for all $i \in [r']$. We set $p_i = \min(1, 2q \cdot u'_i/r)$, where $q = \Theta(r^2 \varepsilon^{-2} \ln(1/\varepsilon))$ is chosen as per Fact 21 with the δ there set to equal a small constant. Then, the number of i for which $2q \cdot u'_i/r \geq 1$ is at most $2q$, since $\sum_i u_i = r$. In addition to these rows being sampled, the expected number of rows sampled is at most

$$\sum_{i=1}^n \frac{2qu'_i}{r} = O\left(\sum_{i=1}^n \frac{qu_i}{r}\right) = O(q),$$

and so by a Markov bound the number of samples is $O(q)$ with arbitrarily large constant probability. The time complexity is $O(\text{nnz}(A') \log(r/\varepsilon) + r^3 \log^2 r + r^2 \log(r/\varepsilon))$.

Now we now have a regression problem on an $O(q) \times r$ matrix SA . Using the Fast Johnson-Lindenstrauss transform, this can be solved in $O(qr \log q + r^3 \varepsilon^{-1} \log r)$ time, see [39, 24]. The success probability is at least $9/10$. This is $O(r^3 \varepsilon^{-2} \log(r/\varepsilon) \log(1/\varepsilon))$ time.

Hence, the overall time complexity is as claimed, and the success probability is at least $2/3$. \blacksquare

7 Low Rank Approximation

While theorems Theorem 11 and Theorem 17 are stated in terms of specific constant probability of success, they can be re-stated and proven so that the failure probabilities are arbitrarily small, but still constant. In the following we'll assume that adjustments have been done, so that the sum of a fixed number of such failure probabilities is at most $1/5$.

7.1 Preliminaries

We collect a few standard lemmas and facts in this subsection which we need for low rank approximation.

Lemma 22 (Approximate Matrix Multiplication) *For A and B matrices with n rows, and given $\epsilon > 0$, there is $t = \Theta(\epsilon^{-2})$, so that for a $t \times n$ generalized sparse embedding matrix S , or $t \times n$ fast JL matrix, or subsampled randomized Hadamard matrix,*

$$\Pr[\|A^\top S^\top SB - A^\top B\|_F^2 < \epsilon^2 \|A\|_F^2 \|B\|_F^2] \geq \delta,$$

for any fixed $\delta > 0$.

Proof: For such a matrix with parameters k and v , first suppose $v = 1$, so that S is the embedding matrix of §2. Let $X = A^\top S^\top SB - AB$. Then $X_{i,j} = A_i^\top S^\top SB_j^\top - A_i^\top B_j$, where A_i is the i -th column of A and B_j is the j -th column of B . Thorup and Zhang [40] have shown that $\mathbf{E}[X_{i,j}] = 0$ and $\mathbf{Var}[X_{i,j}] = O(1/t) \cdot \|A_i\|_2^2 \|B_j\|_2^2$. Consequently, $\mathbf{E}[X_{i,j}^2] = \mathbf{Var}[X_{i,j}] = O(1/t) \cdot \|A_i\|_2^2 \|B_j\|_2^2$, from which for an appropriate $t = \Theta(\epsilon^{-2})$, the lemma follows by Chebyshev's inequality. For $v > 1$, $X_{i,j} = \frac{v}{t} \sum_{i \in [t/v]} \hat{X}_{i,j}$, see (6), so that

$$\mathbf{Var}[X_{i,j}] = \frac{v^2}{t^2} \sum_i \mathbf{Var}[\hat{X}_{i,j}] \leq \frac{v}{t^2} \|A_i\|_2^2 \|B_j\|_2^2 \leq \frac{1}{t} \|A_i\|_2^2 \|B_j\|_2^2,$$

and similarly the lemma follows for the sparse embedding matrices. The result for fast JL matrices was shown by Sarlós[39], and for subsampled Hadamard by Drineas et al.[24]; also the claim follows from norm-preserving properties of these transforms, see [28]. \blacksquare

Fact 23 *Given $n \times d$ matrix A of rank $k \leq n^{1/2-\gamma}$ for $\gamma > 0$, and $\epsilon > 0$, an $m \times n$ fast JL matrix Π with $m = \Theta(k/\epsilon^2)$ is a subspace embedding for A with failure probability at most δ , for any fixed $\delta > 0$, and requires $O(nd \log n)$ time to apply to A .*

A similar fact holds for subsampled Hadamard transforms.

Fact 24 (Normal Equations) Given $n \times d$ matrix C , and $n \times d'$ matrix D consider the problem

$$\min_{X \in \mathbb{R}^{d \times d'}} \|CX - D\|_F^2.$$

The solution to this problem is $X^* = C^-D$, where C^- is the Moore-Penrose inverse of C . Moreover, $C^\top(CX^* - D) = 0$, and so if c is any vector in the column space of C , then $c^\top(CX^* - D) = 0$.

Fact 25 (Pythagorean Theorem) If C and D matrices with the same number of rows and columns, then $C^\top D = 0$ implies $\|C + D\|_F^2 = \|C\|_F^2 + \|D\|_F^2$.

7.2 An Intermediate Theorem

The main theorem in this subsection is the following. It is very close to Lemma 1 of [24].

Theorem 26 Suppose A and B are matrices with n rows, and A has rank at most k . Suppose S is a $t \times n$ matrix, and the event occurs that S satisfies Lemma 22 with error parameter $\sqrt{\epsilon/k}$, and also that S is a subspace embedding for A with error parameter $\epsilon_0 \leq 1/\sqrt{2}$. Then if \tilde{X} is the solution to

$$\min_X \|S(AX - B)\|_F^2, \quad (7)$$

and X^* is the solution to

$$\min_X \|AX - B\|_F^2, \quad (8)$$

then

$$\|A\tilde{X} - B\|_F \leq (1 + \epsilon)\|AX^* - B\|_F.$$

Before proving Theorem 26, we will need the following lemma.

Lemma 27 For S, A, B, X^* and \tilde{X} as in Theorem 26,

$$\|A(\tilde{X} - X^*)\|_F \leq 2\sqrt{\epsilon}\|B - AX^*\|_F.$$

Proof: Let $A = U\Sigma V^\top$ denote the singular value decomposition of A . Since A has rank at most k , we can consider U and V to have at most k columns. Observe that

$$\|A(\tilde{X} - X^*)\|_F = \|U\Sigma V^\top(\tilde{X} - X^*)\|_F = \|U^\top U\Sigma V^\top(\tilde{X} - X^*)\|_F = \|\beta\|_F,$$

where $\beta := U^\top A(\tilde{X} - X^*)$. By Fact 24, we have $A^\top S^\top S(A\tilde{X} - B) = 0$. Since $V^\top \Sigma$ has full column rank, the normal equations imply

$$U^\top S^\top S(A\tilde{X} - B) = 0. \quad (9)$$

To bound $\|\beta\|_F$, we bound $\|U^\top S^\top S U \beta\|_F$, and then show that this implies that $\|\beta\|_F$ is small. Using that $U^\top U = I$ and (9), we have

$$U^\top S^\top S U \beta = U^\top S^\top S U U^\top A(\tilde{X} - X^*) = U^\top S^\top S A(\tilde{X} - X^*) + U^\top S^\top S(B - A\tilde{X}) = U^\top S^\top S(B - AX^*).$$

Using the hypothesis of the theorem,

$$\|U^\top S^\top S U \beta\|_F = \|U^\top S^\top S(B - AX^*)\|_F \leq \sqrt{\epsilon/k}\|U\|_F\|B - AX^*\|_F \leq \sqrt{\epsilon}\|B - AX^*\|_F.$$

To show that this bound implies that $\|\beta\|_F$ is small, we use the property of any conforming matrices C and D , that $\|CD\|_F \leq \|C\|_2\|D\|_F$, obtaining

$$\|\beta\|_F \leq \|U^\top S^\top S U \beta\|_F + \|U^\top S^\top S U \beta - \beta\|_F \leq \sqrt{\epsilon}\|B - AX^*\|_F + \|U^\top S^\top S U - I\|_2\|\beta\|_F.$$

By the hypothesis of the theorem, $\|S U x\|^2 = (1 \pm \epsilon_0)\|x\|^2$ for all x , so that $U^\top S^\top S U - I$ has eigenvalues bounded in magnitude by ϵ_0^2 , which implies singular values with the same bound, so that $\|U^\top S^\top S U - I\|_2 \leq \epsilon_0^2$. Thus $\|\beta\|_F \leq \sqrt{\epsilon}\|B - AX^*\|_F + \epsilon_0^2\|\beta\|_F$, or

$$\|\beta\|_F \leq \sqrt{\epsilon}\|B - AX^*\|_F / (1 - \epsilon_0^2) \leq 2\sqrt{\epsilon}\|B - AX^*\|_F,$$

since $\epsilon_0^2 \leq 1/2$. This bounds $\|\beta\|_F$, and so proves the lemma. ■

Proof of Theorem 26: Let $U\Sigma V^\top$ denote the SVD of A . By Fact 24, and since U and A have the same columnspace,

$$U^\top(AX^* - B) = A^\top(AX^* - B) = 0. \quad (10)$$

This and Fact 25, the Pythagorean Theorem, imply

$$\|A\tilde{X} - B\|_F^2 = \|AX^* - B\|_F^2 + \|A(\tilde{X} - X^*)\|_F^2, \quad (11)$$

which with Lemma 27, implies that

$$\|A\tilde{X} - B\|_F \leq (1 + 2\epsilon)\|AX^* - B\|_F,$$

using $\sqrt{1 + 4\epsilon} \leq 1 + 2\epsilon$. Adjusting and renaming ϵ , the theorem follows. \blacksquare

We use the following observations to analyze the composition of our sparse embedding matrices with fast JL matrices.

Fact 28 *If $S \in \mathbb{R}^{t \times n}$ approximates matrix products and is a subspace embedding with error ϵ and failure probability δ_S , and $\Pi \in \mathbb{R}^{i \times t}$ approximates matrix products with error ϵ and failure probability δ_Π , then ΠS approximates matrix products with error $O(\epsilon)$ and failure probability at most $\delta_S + \delta_\Pi$.*

Proof: This follows from two applications of Lemma 22, together with the observation that $\|SAx\| = (1 \pm \epsilon)\|Ax\|$ for basis vectors x implies that $\|SA\|_F = (1 \pm \epsilon)\|A\|_F$. \blacksquare

Fact 29 *If $S \in \mathbb{R}^{t \times n}$ is a subspace embedding with error ϵ and failure probability δ_S , and $\Pi \in \mathbb{R}^{i \times t}$ is a subspace embedding with error ϵ and failure probability δ_Π , then ΠS is a subspace embedding with error $O(\epsilon)$ and failure probability at most $\delta_S + \delta_\Pi$.*

7.3 Putting it All Together

Let $\Delta_k \equiv \|A - A_k\|_F$, where A_k is the best rank- k approximation to A . The following theorem is the same as Theorem 4.7 of [8] except that we use our sparse embedding matrices instead of the dense sign matrices used in [8].

Theorem 30 *Let A be an $n \times n$ matrix, and $k \in [n]$ with $k < n^{1/2-\gamma}$ with $\gamma > 0$. Let \hat{R} and \hat{S} be independent $t_v \times n$ and $t'_v \times n$ generalized sparse embedding matrices, respectively, for parameters $v = 1$ with $t = t_1 = \Theta(k/\epsilon + k^4 \log^2 k)$ and $t'_1 = \Theta((k/\epsilon)^4 \log^2(k/\epsilon))$ or $v_{\text{JL}} = \Theta(\log r)$ with $t_{v_{\text{JL}}} = \Theta(k/\epsilon + k^2)$ and $t'_{v_{\text{JL}}} = \Theta(k\epsilon^{-4} \log(k/\epsilon)(k + \log(1/\epsilon)))$. Let Π_S be a fast JL operator from $\mathbb{R}^{t'_v}$ to $\mathbb{R}^{O(k/\epsilon^2)}$, and let Π_R be a fast JL operator from \mathbb{R}^{t_v} to $\mathbb{R}^{O(k/\epsilon)}$. Let $S \equiv \Pi_S \hat{S}$, and $R \equiv \Pi_R \hat{R}$. Let*

$$\hat{A} = AR^\top (SAR^\top)^- SA.$$

Then

$$\Pr[\|A - \hat{A}\|_F \leq (1 + \epsilon)\|A - A_k\|_F] \geq 1 - \delta,$$

for any fixed $\delta > 0$. Moreover, AR^\top and ZA and $(SAR^\top)^-$ can be computed in $O(v^2 \text{nnz}(A) + (n + t_v)t'_v \log n)$ time. Alternatively, a similar decomposition can be computed, without fast JL matrices, in $O(v^2 \text{nnz}(A) + \text{poly}(k/\epsilon))$ time, with factors of dimension $n \times m$, $m \times m'$, and $m' \times n$, where $m, m' = O(\text{poly}(k/\epsilon))$. For $k > n^{1/2-\gamma}$, the subsampled randomized Hadamard transforms can replace the fast JL transforms, with $\log^2(k/\epsilon)$ increase in size for the transforms.

Proof: For the given parameters, the event of Theorem 26 occurs for \hat{S} and a rank $O(k/\epsilon)$ matrix with arbitrarily small constant failure probability, since $t'_v = \Omega(k/\epsilon^2)$ suffices for Lemma 22 to apply, and t'_1 suffices for Theorem 11 to apply, while $t'_{v_{\text{JL}}}$ suffices for Theorem 17 to apply, and similarly for Π_S , and for

S using Facts 28 and 29, with a total failure probability that is an arbitrarily small constant. We apply Theorem 26 to S , with k , A , and B of the theorem mapping to k/ϵ , AR^\top , and A , respectively.

The result is that with small fixed failure probability, for \tilde{X} the solution to

$$\min_X \|SAR^\top X - SA\|_F,$$

we have

$$\|AR^\top \tilde{X} - A\|_F \leq (1 + \epsilon) \|AR^\top X^* - A\|_F = (1 + \epsilon) \min_X \|AR^\top X - A\|_F.$$

Similarly the parameters for \hat{R} suffice for Theorem 26 to apply with k , A , B , and t of the theorem mapping to k , A_k^\top , A^\top , and t_v , respectively, and similarly for Π_R and R .

We have, with small fixed failure probability,

$$\|AR^\top X^* - A\|_F \leq (1 + \epsilon) \|A - A_k\|_F = (1 + \epsilon) \Delta_k, \quad (12)$$

because $\tilde{Y} \equiv AR^\top (A_k R^\top)^\top$ has, by Theorem 26,

$$\|\tilde{Y} A_k - A\| \leq (1 + \epsilon) \min_Y \|Y A_k - A\| = (1 + \epsilon) \Delta_k,$$

and so

$$\min_X \|AR^\top X - A\| \leq \|AR^\top (A_k R^\top)^\top A_k - A\| \leq (1 + \epsilon) \Delta_k.$$

Since $\tilde{X} = (\Pi_S SAR^\top)^\top SA$, we have

$$\|AR^\top (SAR^\top)^\top SA - A\|_F = \|AR^\top \tilde{X} - A\|_F \leq (1 + \epsilon) \|AR^\top X^* - A\|_F \leq (1 + \epsilon)^2 \Delta_k,$$

and the first part of the theorem follows, after adjusting ϵ by a constant factor.

For the time complexity, because \hat{R} and \hat{S} are sparse embedding matrices, we can compute $\hat{S}A$ and $A\hat{R}^\top$ in $O(v \text{nnz}(A))$ time, and because Π_R and Π_S are fast JL matrices, we obtain SA and AR^\top in $O(nt'_v \log n)$ time. From this, we can compute SAR^\top in $O(v^2 \text{nnz}(A) + t_v t'_v \log n)$ time, and $(SAR^\top)^\top$ in $O(k^3/\epsilon^4)$ time. \blacksquare

Theorem 31 (Main) *Given an $n \times n$ matrix A , with probability at least $3/5$, we can compute the following triple of matrices (L, D, W) , where L is an $n \times k$ matrix, D is a $k \times k$ diagonal matrix, and W is a $k \times n$ matrix, with the property that*

$$\|A - L \cdot D \cdot W\|_F \leq (1 + \epsilon) \Delta_k,$$

that is, $L \cdot D \cdot W$ is within $(1 + \epsilon)$ of the best rank- k approximation. Furthermore, L , D , and W can be computed in $O(v^2 \text{nnz}(A) + (n + t'_v(\epsilon^2))t'_v(\epsilon^2))$ time, where $t'_1(\epsilon^2) \equiv \Theta((k/\epsilon^2)^4 \log^2(k/\epsilon))$, and $t'_{v_{\text{JL}}}(\epsilon^2) \equiv \Theta(k\epsilon^{-8} \log(k/\epsilon)(k + \log(1/\epsilon)))$.

Proof: By Theorem 30, we can compute AR^\top , $(SAR^\top)^\top$, and SA in the given time, for which with constant probability,

$$\|A - \tilde{A}\|_F \leq (1 + \epsilon) \Delta_k,$$

where $\tilde{A} \equiv AR^\top (SAR^\top)^\top SA$.

By Theorem 4.8 of [8], under these conditions it follows that if U is an orthonormal basis for the column space of AR^\top , then

$$\|A - U[U^\top \tilde{A}]_k\|_F \leq (1 + \sqrt{\epsilon}) \Delta_k,$$

where $[U^\top \tilde{A}]_k$ denotes the best rank- k approximation to $U^\top \tilde{A}$. We replace ϵ by ϵ^2 adjust the $t_v = t_v(\epsilon^2)$ values accordingly, and have

$$\|A - U[U^\top \tilde{A}]_k\|_F \leq (1 + \epsilon) \Delta_k.$$

Notice that $U[U^\top \tilde{A}]_k$ is a matrix of rank k . Given AR^\top , we can compute U in time $O(nk^2/\epsilon^4)$, since AR^\top is $n \times O(k/\epsilon^2)$. We can also compute $U^\top \tilde{A}$ in $O(nk^2/\epsilon^6)$ time by first computing $U^\top AR^\top$, then

$U^\top AR^\top (SAR^\top)^-$, and then $U^\top AR^\top (SAR^\top)^- ZA$. We can compute the best rank- k approximation $[U^\top \tilde{A}]_k$ to $U^\top \tilde{A}$ in time $O(nk^2/\epsilon^4)$ since $U^\top \tilde{A}$ is an $O(k/\epsilon^2) \times n$ matrix. Further, we can compute the SVD $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$ of $[U^\top \tilde{A}]_k$ in $O(nk^2/\epsilon^4)$ time.

Finally, we set $L = U\tilde{U}$, $D = \tilde{\Sigma}$, and $W = \tilde{V}^\top$. It follows that L, D , and W can be computed in $O(v^2 \mathbf{nnz}(a)) + (n + t'_v(\epsilon^2)t'_v(\epsilon^2) + nk^2/\epsilon^4)$ time, where the last summand can be omitted since the previous one dominates. By construction of L, D , and W , the requirements of the theorem are satisfied. This completes the proof. \blacksquare

8 Preliminary Experiments

Some preliminary experiments show that the low-rank approximation technique of Theorem 30 is quite promising, and in practice may perform much better than the general bounds of the theorem. In the experimental version, the matrices tested are $n \times d$, with $d \leq n$ without loss of generality, and the embedding matrices are $t \times n$ and $d \times t^2$ (or the original matrix for large t).

The resulting low-rank approximation was tested for t taking values of the form $\lceil 1.4^z \rceil$, with $t \leq d/4$, and for each such t , taking the ratio R_e of the Frobenius norm of the error with the Frobenius norm of the best rank- k approximation, for all $k \leq t$. The resulting points $(t/k, R_e)$ were generated, for all test matrices, resulting in a set of points P . (Actually, three instantiations of the low-rank approximations were done, generating different embedding matrices, and the results are the middle value of the three.)

The test matrices are from the University of Florida Sparse Matrix Collection, essentially all those with at most 10^5 nonzero entries, and with n at most about 600, because the tests required some slow operations and were done on a laptop. However, despite such restrictions, over 500 matrices were tested, taken from 40 sub-collections, each such sub-collection representing a particular application area.

The results are shown in two plots. Figure 1 shows the k -extremal curves of the point set P , for $k = 1, 3, 5$, with y -coordinates that $\log(R_e)$, using the natural logarithm. Here a point (τ, ρ) on a k -extremal curve implies that for all but k matrices, tests run with $t/k \geq \tau$ yielded ratios $\log(R_e) \leq \rho$. Although the curves are shown for $R_e > 1$, for many matrices $R_e < 1$ for large enough t/k .

In Figure 2, only the 5-extremal curve is shown, for $t/k > 4$ (and the y coordinate is R_e , not its log). We see for $t/k \geq 4$, the error ratio for all but 5 matrices is no more than 1.21, and this decreases to less than 1.01 for t/k about 11.

References

- [1] Dimitris Achlioptas, Amos Fiat, Anna R. Karlin, and Frank McSherry. Web search via hub synthesis. In *FOCS*, pages 500–509, 2001.
- [2] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *COLT*, pages 458–469, 2005.
- [3] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2), 2007.
- [4] Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *APPROX-RANDOM*, pages 272–279, 2006.
- [5] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *STOC*, pages 619–626, 2001.
- [6] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.
- [7] Ho Yee Cheung, Tsz Chiu Kwok, and Lap Chi Lau. Fast matrix rank algorithms and applications. In *STOC*, pages 549–562, 2012.

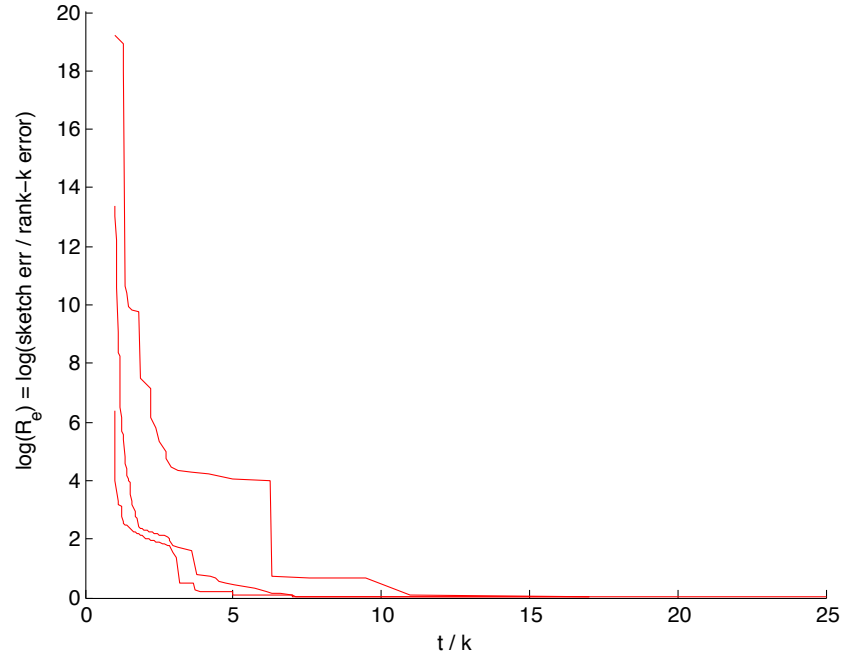


Figure 1: k -extremal curves of performance for all matrices, $k \in \{1, 3, 5\}$

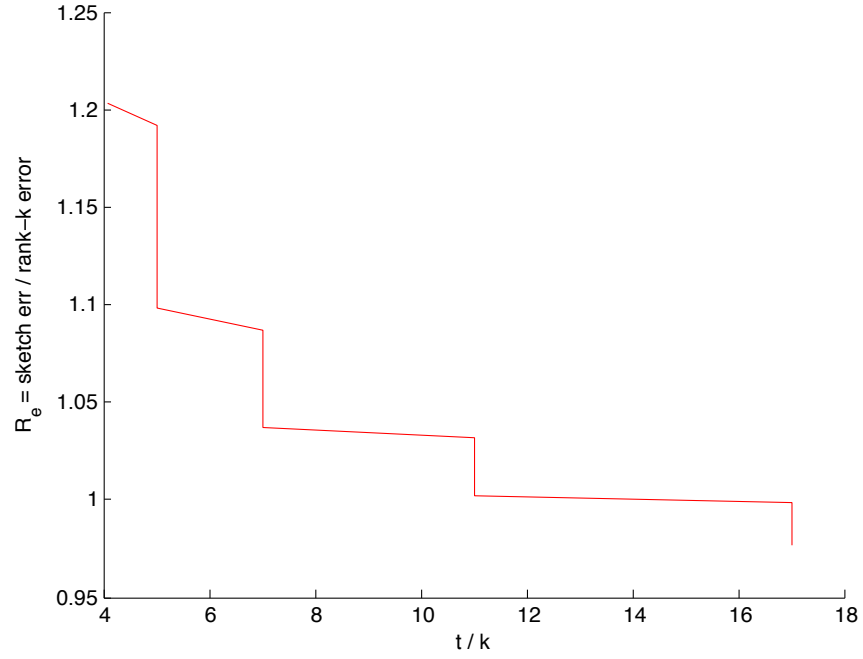


Figure 2: 5-extremal curve with $t/k \geq 4$

- [8] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.
- [9] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.
- [10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *STOC*, pages 341–350, 2010.
- [11] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *SODA*, pages 1117–1126, 2006.
- [12] Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *APPROX-RANDOM*, pages 292–303, 2006.
- [13] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.
- [14] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.
- [15] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1):158–183, 2006.
- [16] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36(1):184–206, 2006.
- [17] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *STOC*, pages 82–90, 2002.
- [18] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *CoRR*, abs/1109.3843, 2011.
- [19] Petros Drineas, Michael Mahoney, Malik Magdon-Ismail, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *ICML*, 2012.
- [20] Petros Drineas and Michael W. Mahoney. Approximating a Gram matrix for improved kernel-based learning. In *COLT*, pages 323–337, 2005.
- [21] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *SODA*, pages 1127–1136, 2006.
- [22] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, pages 316–326, 2006.
- [23] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *ESA*, pages 304–314, 2006.
- [24] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *CoRR*, abs/0710.1435, 2007.
- [25] Alan M. Frieze, Ravi Kannan, and Santosh Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004.
- [26] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *ArXiv e-prints*, September 2009.
- [27] D.L. Hanson and F.T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

- [28] Daniel M. Kane and Jelani Nelson. A sparser Johnson-Lindenstrauss transform. *CoRR*, abs/1012.1577, 2010.
- [29] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012.
- [30] Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *STOC*, pages 745–754, 2011.
- [31] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.
- [32] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
- [33] Avner Magen and Anastasios Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *SODA*, pages 1422–1436, 2011.
- [34] Frank McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537, 2001.
- [35] Jelani Nelson and David P. Woodruff. Fast Manhattan sketches in data streams. In *PODS*, pages 99–110, 2010.
- [36] Nam H. Nguyen, Thong T. Do, and Trac D. Tran. A fast and efficient algorithm for low-rank approximation of a matrix. In *STOC*, pages 215–224, 2009.
- [37] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci.*, 61(2):217–235, 2000.
- [38] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4), 2007.
- [39] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.
- [40] Mikkel Thorup and Yin Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA*, pages 615–624, 2004.
- [41] Lloyd N. Trefethen and David Bau. *Numerical linear algebra*. SIAM, 1997.
- [42] Anastasios Zouzias and Nikolaos M. Freris. Randomized extended Kaczmarz for solving least-squares. *CoRR*, abs/1205.5770, 2012.